ABSTRACT

The present invention relates to a system and methodology to facilitate extraction of information from a large unstructured corpora such as from the World Wide Web and/or other unstructured sources. Information in the form of answers to questions can be automatically composed from such sources via probabilistic models and cost-benefit analyses to guide resource-intensive information-extraction procedures employed by a knowledge-based question answering system. The analyses can leverage predictions of the ultimate quality of answers generated by the system provided by Bayesian or other statistical models. Such predictions, when coupled with a utility model can provide the system with the ability to make decisions about the number of queries issued to a search engine (or engines), given the cost of queries and the expected value of query results in refining an ultimate answer. Given a preference model, information extraction actions can be taken with the highest expected utility. In this manner, the accuracy of answers to questions can be balanced with the cost of information extraction and analysis to compose the answers.